DOCUMENT RESUME

ED 233 073 TM 830 523

AUTHOR Johnson, E. Marcia; Wolfe, Richard G.
TITLE Implementing and Evaluating a Research Data Archive.
PUB DATE Apr 83
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Archives; Computer Oriented Programs; Data; Information Needs; *Information Networks; *Information Processing; *Research Needs; Research Problems; *Statistical Data
IDENTIFIERS *Ontario Institute for Studies in Education

ABSTRACT

This paper addresses some of the problems of data access and exchange by discussing what an educational research institution should do about data archiving and data facilities. The work undertaken at the Ontario Institute for Studies in Education is described. Generally, efforts in the past have been given to computerization of bibliographic information and to computerization of data analysis facilities. It is now time to turn to the problems of the research data resource itself--to data development, data archiving, and to computerization of data (and documentation) access and exchange. A first step is for research institutions, government agencies, university departments, and private centers to establish a local data archive and computing laboratory, dedicating as much attention to this as has been given to print collections and to analysis software. Institutions and contracting agencies should develop and adhere to reasonable guidelines about data preservation on the one hand, and data ownership on the other. The computerization of data access and exchange between institutions requires good arrangements at each institution and standardized facilities for computer networks. Another need is for a centralized archive function. (Author/PN)

Implementing and Evaluating
A Research Data Archive

E. Marcia Johnson & Richard G. Wolfe

Ontario Institute for Studies in Education

A paper presented at the meetings of the American Educational Research Association, Montreal, Quebec, April 11-15, 1983.

ABSTRACT

Everyone is talking about the "information explosion". In most fields of research, more papers have been published since World War II than in all years previously (Martin, 1981). The explosion in research publication is fueled by an even greater explosion in research data - one that exceeds our capacity, at least in educational science, to perform adequate analysis and synthesis of findings. As a consequence, we are urged to carry out secondary analysis (Burstein, 1977), and to carry out meta-analyses integrating the results of many studies (Glass, 1976). Secondary analysis and integrative re-analysis usually require our locating and accessing the original research data files in their machine-readable form. Any educational researcher who has attempted this can testify that our collegial and institutional mechanisms for data access and exchange are primitive! The purpose of this paper is to address some of these problems by discussing what an educational research institution, for example, a university department, should do about data archiving and data facilities. The work undertaken at the Ontario Institute for Studies in Education (OISE) will be described.

1.0 TECHNOLOGICAL CHANGE

With the development and proliferation of large-scale computer systems and cheap data storage in the 1960's and 1970's came the development of computerized bibliographic information retrieval systems and large university and government data archives.

In 1969, as an outgrowth of the aerospace industry, the Lockheed Missile and Space Company Inc. launched its commercial service (DIALOG) with a single database called ERIC. Researchers can now obtain the latest

reports of educational research projects by manually searching ERIC microfiche reports in their local university library; by requesting on-line library searches of the ERIC database; or, more recently, if they have a computer terminal and modem in their office, by conducting their own on-line searches of ERIC. This ability of educational researchers to access quickly and easily educational research findings in ERIC and other databases, such as Psychological Abstracts, has changed educational research. Information on a variety of topics, new methodological techniques, new theoretical perspectives can all be located outside of the researcher's personal communications network.

Coupled with the ability of researchers to access research findings has been the development of sophisticated facilities for large-scale data analysis. The most well-known of the statistical analysis packages are SPSS, OSIRIS, and PSTAT. More recent and powerful facilities have been developed and made available through SAS and SIR. Not only have researchers been able to take advantage of the increased computing power available since the 1960's, but, more importantly, the statistical packages have forced some degree of standardization on research methodologies.

Concurrently, developments in the telecommunications and software industries have led to the emergence of long haul networks (including those using satellites) and local area networks. What this means in real terms is an increased capability to move data from one place to another at a relatively small cost. But more than data can be effectively shared. It is now technologically possible to share computing power, expensive peripheral devices such as laser printers, and expensive software and processing facilities. One example, in fact, of an educational network is

EDUNET through which computing and data processing facilities are shared by many researchers at diverse sites.

One of the more common complaints about network technology, however, has been the lack of standardization between existing technologies and the accompanying maze of access procedures. There has been considerable research into methods of improving telecommunication capabilities, including, in Canada, the iNET trials. This network system was created in response to the need to make information services more universally accessible through a single communications link that uses simple access codes. The term iNET refers "to the intelligent network functions inherent in the service. iNET is an electronic directory and gateway, designed to simplify the process of gathering information from computer-based sources through a single point of access" (Farhood, 1982). In other words, it is not necessary that a person wishing to use data from several different sources be forced to remember several different network access codes. There is one simple code for accessing iNET, and the network handles the rest of the work. Furthermore, since the service operates through the public network, Datapac, users can access databases available through the United States and other countries' packet switching networks. While this technology is still in the developmental stage, it provides an indication of what could be generally available.

However, in spite of these advancements, we would have to conclude that data access and exchange is still in a primitive state. In fact, the first problems a secondary analyst encounters is locating data for reanalysis (Powell, 1977; Boruch, 1981). Often, the second problem is obtaining permission and arranging access or transfer of data. While the

telecommunications facilities are in place, the mechanisms for informing people of the existence of new data sets are often not effective for reaching a large research community, and the political (administrative, budgetary, etc.) mechanisms provide obstacles rather than facilitation. The upshot is that data location and access takes weeks or months rather than seconds! What, then should be considered as prerequisites to the electronic transfer of data analysis and research findings? How do we organize data to maximize its use?

2.0 SUMMARY OF DATA ARCHIVING

One method of organizing data to facilitate electronic transfer would be to establish a network of educational research institutions each having a data archive at its node. A data archive is like a library, but the contents are machine-readable data files (MRDF), together with the documentation of the variables, formats, sampling frame. The data files are "clean" -- that is, free of obvious coding errors and reconciled with the documentation.

The data archiving profession emerged in university and government institutions in response to the proliferation of MRDF. The major components of data archiving include not only documenting and abstracting MRDF, but also cataloguing and indexing the files. Dodd (1979) stated that "the development of information technology and the ability to produce data have progressed much more rapidly than our capacity to organize, classify, and reference its availability." Since the publication of Dodd's article, several important changes have occurred in the standardization of procedures for describing data files. These include the incorporation of

descriptive cataloguing procedures for MRDF in the second edition of the Anglo American Cataloguing Rules and the 1982 publication of Dodd's book Cataloguing Machine-Readable Data Files: An Interpretative Manual.

The role of the data archivist includes another important element; namely, the location and selection of appropriate data files. In fact, the issue of selecting data for inclusion in an archive has been discussed in workshops of the IASSIST annual conference (1980) and is represented in the literature (Robbin, 1977). Many granting and contracting agencies, especially governmental ones, have been including stipulations that data be given to them in machine-readable form (maybe even in SAS or SPSS files). While this is an important prerequisite step, it does not begin to answer the questions of selecting data for real archiving. For that will require (a) substantial additional investment in data development, cleaning, redocumenting, cataloguing, etc; and consequently (b) very careful determination of which data files are worth archiving.

Ultimately, the success of the work of data archivists will come to depend on the research community environment in which it is placed. There are serious gaps between the usual computing facilities at educational research centres and what is technologically feasible and desireable. Three areas seem particularly serious.

First, older types of computer systems in use at many archive sites are not adequate. The predominent method of data storage is magnetic tape, and as any researcher using this storage medium can verify, tapes are slow due to the sequential access method necessary to locate desired data files; they are prone to errors; and the recording medium itself is not stable after five or more years. While the price of disk storage has lowered

dramatically over the past 5 or 10 years, and tape storage has not, we have not adjusted our capabilities and practices. Furthermore, in the field of educational research, non-numeric data are becoming more popular for qualitative analysis. This includes, for example, the analysis of textual materials. Trying to perform these analyses on technologically out-dated systems is virtually impossible.

Second, few university computing centres have on-line data management facilities to allow researchers to cross-reference materials or relate findings from diverse disciplines. Standards for on-line data dictionaries do not exist. At the least, on-line directories of research data stored in the archive should be available for easy location of appropriate studies.

Third, our computing facilities (with some notable exceptions, such as EDUNET) have remained isolated. There is, for example, inter-institutional access and exchange of data taking place electronically.

These difficulties, then, present formidable barriers to researchers wanting to reach across data networks. At the Ontario Institute for Studies in Education some new techniques were considered in the planning and implementation of a research data archive.

3.0 THE OISE DATA ARCHIVE AND LABORATORY

Since its establishment in 1965, OISE has been a major producer, processor, and user of educational research data. Literally thousands of studies have involved testing students, questioning teachers, interviewing parents, and so on. However, until recently, there had been no systematic effort within OISE to organize and preserve data, and there had been little

encouragement to carry out secondary analysis. In 1979, under a Ministry of Education contract, and in cooperation with the Institute for Behavioural Research, York University (Atkinson & Wolfe, 1980), an educational research data archive and laboratory was established at OISE. Examination of the histories of other social science archives, some of them extinct, guided the design of the data archive and computing laboratory. It was recognized that simply generating new data resources would not be sufficient to ensure increased secondary analysis. Data resources had to be made available to potential secondary analysts in a context which would facilitate (1) the location of appropriate data, (2) the design of analysis strategies, and (3) the implementation of analysis and interpretation of results. It was for these purposes that the data laboratory component was incorporated into the Computing Services Group (CSG) at OISE.

The major components of the existing data archive facility can be summarized as follows:

- Over thirty files of educational and historical importance have been prepared for secondary analysis.
- A collection of over thirty-five small instructional data files intended for use in methodology courses has been established. These files have often been abstracted from the major files, but also we have attempted to capture data from thesis research, textbook examples, published papers, etc.
- A comprehensive package of well-documented data analysis programs are available for use.
- A series of Bulletins describing all aspects of the archive facility has been published and is distributed regularly to academic

departments.

- Statistical consultants and systems analysts within the CSG are available to provide guidance with design strategies and with the implementation and interpretation of statistical analyses.
- Extensive contacts are maintained with other data archives and data clearinghouses. Catalogues of their holdings are kept in the archive office, and researchers are invited to make use of the facilities at other institutions by channelling requests through the data archivist.
- Contact with the user community is frequent.

Nesvold (1976) addressed some of the problems encountered in training scholars in research methods, data analysis and "the analytical approaches to the study of social problems". These problems included access to small, interesting data files as well as access to clear, concise documentation of the data. We were motivated by considerations like these in the initial design and implementation of the instructional dataset collection. Instructions detailing how to access the files from the archive, and complete file documentation are stored on-line. Continuing data development has included the conducting of case studies of the individual methodology courses several times during the year. Information on the appropriateness of existing instructional files is requested from course instructors and suggestions for new types of datasets to include in the collection are solicited.

Research projects (new and continuing) and student thesis research are monitored regularly. Lists of the research projects are obtained from the OISE department of Field Services and Research and lists of graduate theses are obtained from the academic departments. Using these listings as

guides, we personally contact the principal investigators of data-producing projects as well as recent graduates and invite them to donate their data to the archive.

In addition to personal contacts, the evaluation of the archive and laboratory as a service institution is based on measurement of the usage patterns. As part of the operation of the OISE computer, a mechanism has been established for maintaining a log of all accesses of archive files. This measurement process has facilitated the monitoring of such information as purpose and frequency of access, type of user, and program used with the data.

Three years after completing the initial data archive and laboratory at OISE we are favourably impressed with the usage for instructional applications, but less impressed with the number of major research applications. There have been a few theses which have made use of major data files, but this kind of use of the archive holdings could be greatly expanded.

Research into utilization has proven effective as a means of locating obvious weaknesses or gaps in both the archive collection and service. The information obtained from the evaluations has allowed us to tailor the instructional data files and the acquisition process for the major files to the needs of the OISE community. Needs for specific types of data can be identified.

The principal of easy access to data cannot be emphasized too strongly. All of the instructional datasets as well as corresponding documentation are kept on-line on the OISE computer so that access is fast

and simple. This, combined with a team approach to providing archiving, statistical and computing (programming) services has been a significant feature of the archive.

4.0 DATA EXCHANGE NETWORKS

We have found that frequent personal contact with the user community has been helpful in obtaining data, but moreover has raised the awareness of educational researchers in the area of data archiving and secondary analysis. While we have had some success dealing with the problems of data access, data exchange remains a major hurdle. However, an imminent change in our computing facilities will provide new possibilities.

Our provincial government has formed the educational computing network of Ontario (ECNO) in an attempt to unify and share data processing operations and development costs among Ontario school boards. The network is based on VAX minicomputers, and, consequently, most school boards can afford to buy the hardware. (There are four different sizes (and prices) of VAX, and all machines in the family are software compatible). OISE is just purchasing two VAX 11/750 computers, and negociations to join ECNO are in progress.

The network will first of all facilitate communication among educational researchers and practitioners and will speed exchange of data and programs. The immediate emphasis is on administrative software and data. But there are two specific software capabilities that have the potential for a great influence on our work. First, DECNET is a networking package that allows users of one computer, say in a school board office, to

access and analyse data in a file on another computer, say at OISE, with practically no extra specifications or arrangements. (Of course, this facility needs to be established with appropriate file security). The second package, DATATRIEVE, is a data dictionary, and file management and reporting program that stores general definitions of data variables, records, and files. DATATRIEVE facilitates access and analysis of data across the network. In light of our experience with the OISE internal data archive and laboratory, these facilities are ideal: DECNET automates the inter-system access, eliminating all layers of bureaucracy. DATATRIEVE automates access to information about data (documentation). This parallels what we have found so successful in our instructional database.

It should be noted that most universities around North America have one or more VAX computers. What are the possibilities for larger-scale networks?

5.0 FUTURE OF DATA EXCHANGE: OWNERSHIP

The possibility of improved data exchange capabilities leads to the issue of ownership of data. Data archivists have always been acutely aware of this problem, but with the advent of distributed processing environments, the difficulties are magnified. In fact, there exists a long academic tradition of personal ownership of data from which we should try to break away. Boruch (1981) stated that

> a recent argument over the integrity of evidence on the effects of diabetes control methods bore more similarity to a war than to a scholarly exchange of views....The dissemination of data for reanalysis was blocked, in Forsham vs. Califano, in the interests of territorial rights - the proprietary rights of the investigator - and probably also the wish not to succor the enemy.

This attitude, while understandable, reduces the attractiveness of secondary analysis to researchers.

More than ownership of the data is at stake however. The issue of the privacy and protection of original respondants must be considered. While it is possible for the primary researcher to guarantee these rights, preserving the anonymity of research participants may be less possible once the data have been released for secondary analysis. Data archivists must be particularly sensitive to the question of respondants' privacy rights. They can use a number of procedures to modify data sets and thus reduce the probability of a sample being reconstructed.

Cross-border data transfer, already a problematic area, would become more of one as data exchange capabilities improve. Although this issue is too complex to be dealt with adequately in this paper, readers may find the European experience interesting. Their bibliographic information retrieval developments through DIANE (Direct Information Access Network for Europe) have been interesting and noteworthy.

6.0 RECOMMENDATIONS AND CONCLUSIONS

Generally, efforts in the past have been given to computerization of bibliographic information and to computerization of data analysis facilities. These efforts have been very successful. It is now time to turn our efforts to the problems of the research data resource itself--to data development, data archiving, and to computerization of data (and documentation) access and exchange. A first step is for research institutions, government agencies, university departments, and private

centres to put their individual houses in order.

(A) Every institution should establish a local data archive and computing laboratory, dedicating as much attention to this as has been given to print collections and to analysis software. This should include: major files, instructional files, catalogues, information and liaison to other archives. There should be an on-going program of selecting data for development.

(B) Institutions and contracting agencies should develop and adhere to reasonable guidelines about data preservation, on the one hand, and data ownership on the other. It must be recognized that data development, cataloguing, etc., has a definite cost.

Our discussion has been concerned with how an individual institution might deal with the problem of data management. What about the larger educational research community? We have raised some flags concerning aspects of data organization that are certain to become more important. One is the computerization of data access and exchange between institutions. This requires good arrangements at each institution and standardized facilities for computer networks. A second is the need for some centralized archive function. Perhaps it is time for a major feasibility analysis of data resource management in educational research.

BIBLIOGRAPHY

Atkinson, T.H. & Wolfe, R.G. Data archive: Ontario educational research. Final report. Ontario Ministry of Education project number 244-OISE. Ontario Institute for Studies in Education, 1980.

Boruch, R.F., Wortman, P.M., & Cordray, D.S. (eds). Reanalyzing program evaluations, San Francisco: Jossey-Bass Inc., 1981.

Burstein, L. Secondary analysis: an important resource for educational research and evaluation. Educational Researcher, 1978, 7(5), 9-12.

Dodd, S. Bibliographic references for numeric social sciences data files: suggested guidelines. Journal of the American Society for Information Science, 1979, 77-82.

Dodd, S. Cataloguing machine-readable data files: an interpretative manual, 1982.

Farhood, L. iNET provides easy access to computer-based services. CIPS Review, 6(6), 1982.

Glass, G.V. Primary, secondary, and meta-analysis of research. Educational Researcher, 1976, 5(10), 3-8.

Martin, J. Telematic society: a challenge for tomorrow, Englewood Cliffs: Prentice-Hall, 1981.

Nesvold, B.A. Instructional applications of data archive resources. American Behavioral Scientist, 1976, 19(4), 455-466.

Powell, M. Necessary steps to insure availability of data for secondary analysis. Paper presented at the Annual Meeting of the American Educational Research Association, New York, April 1977.

Robbin, A. The pre-acquisition process: a strategy for locating and acquiring machine-readable data. Drexel Library Quarterly, 1977, 13(1), 21-42.